

AD-A125 759

QUEUEING MODELS FOR DESIGNING DIGITAL COMMUNICATION  
SATELLITE SYSTEMS(U) STANFORD UNIV CA  
F S HILLIER ET AL, DEC 82 TR-209 N00014-75-C-0561

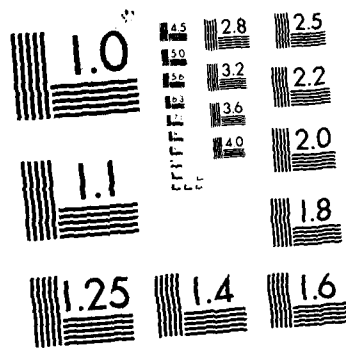
1/1

UNCLASSIFIED

F/G 17/2

NL


END  
FILMED  
DTC



MICROCOPY RESOLUTION TEST CHART  
NATIONAL BUREAU OF STANDARDS-1963-A

Accession For	
NTIS GRA&I	<input checked="checked" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A	

QUEUEING MODELS FOR DESIGNING  
DIGITAL COMMUNICATION SATELLITE SYSTEMS

by

Frederick S. Hillier  
and  
Bijan Jabbari

TECHNICAL REPORT NO. 209

December 1982

SUPPORTED UNDER CONTRACT N00014-75-C-0561 (NR-047-200)  
WITH THE OFFICE OF NAVAL RESEARCH

Gerald J. Lieberman, Project Director

Reproduction in Whole or in Part is Permitted  
for any Purpose of the United States Government  
Approved for public release; distribution unlimited

DEPARTMENT OF OPERATIONS RESEARCH  
AND  
DEPARTMENT OF STATISTICS  
STANFORD UNIVERSITY  
STANFORD, CALIFORNIA

QUEUEING MODELS FOR DESIGNING  
DIGITAL COMMUNICATION SATELLITE SYSTEMS

by

Frederick S. Hillier and Bijan Jabbari

Abstract

Satellites have an enormous potential for providing efficient communication links between many widely scattered ground stations. By interpreting the messages as customers in a queueing system, an unusual type of queueing model can be formulated to describe this process. The expected waiting time of the messages in the queue can then be derived for different configurations in order to guide the design of the system.

## GLOSSARY

- $\tau$ : Random variable representing the interarrival times between messages. See Section 2.1.
- $\lambda$ : Arrival rate of messages ( $=1/E(\tau)$ ).
- $P_n(t)$ : Probability that  $n$  messages arrive during time  $t$  (assumed to be a Poisson distribution).
- $Y$ : Random variable representing the message length (assumed to have an exponential distribution). See Section 2.2.
- $1/\mu$ : Expected mean message length ( $=E(Y)$ ).
- $P$ : Packet size expressed as the number of bits.
- $L$ : Random variable representing the number of packets in a message.
- $T$ : Frame time. See Section 3.1.
- $K$ : Total number of slots in a frame.
- $C$ : Channel transmission rate.
- $a$ : Traffic intensity in erlangs. See Section 3.2.
- $S$ : Number of slots assigned to a traffic source.
- $\rho$ : Normalized traffic intensity.
- $M$ : Buffer capacity (in number of packets). See Section 4.1.
- $\bar{P}$ : Transition probability matrix for the imbedded Markov chain. See Section 4.2.
- $\pi_j$ : Steady state probability of  $j$  messages in the system.
- $R$ : Random variable representing the number of messages rejected (blocked) in one frame.
- $P_B$ : Probability that a message will be blocked.
- $\omega$ : Random variable representing the waiting time in the buffer by a randomly selected arriving message.

$W_q$ : Expected delay in the buffer incurred by messages which are not blocked.

$E_{in}$ : Event that a random arrival finds  $i$  messages left behind in the buffer from the last epoch and is one of  $n$  arriving messages in this frame.

$P(E_{in})$ : Probability that the event  $E_{in}$  occurs for a random arrival.

$\bar{n}$ : Number of new arrivals that wait (not blocked) for Case 2.

$g(\ell)$ : Probability distribution of  $L$ . See Section 4.3.

$R_j$ : Random variable representing the number of messages rejected (blocked) in one frame given that  $j$  packets were left behind in the buffer at the beginning of the frame.

$G^{(i)}(t)$ : Cumulative distribution function of the number of packets in  $i$  messages.

$N$ : Random variable representing the number of messages arriving in a frame.

$E_{ijnpq}$ : Event that a random arrival meets four conditions defined in Section 4.3.

$P(E_{ijnpq})$ : Probability that the event  $E_{ijnpq}$  occurs for a random arrival.

$r$ :  $\min \{p + q, M - i\}$  for the event  $E_{ijnpq}$ .

$g^{(j-1)}(p)$ : Probability that the first  $(j-1)$  messages contain  $p$  packets.

# QUEUEING MODELS FOR DESIGNING DIGITAL COMMUNICATION SATELLITE SYSTEMS<sup>1</sup>

by

Frederick S. Hillier and Bijan Jabbari

## 1. Introduction

Digital communication satellites have an enormous potential for providing efficient and reliable communication links between many distant points. However, the full realization of this potential requires improved techniques for designing digital satellite networks. Of particular importance is the channel architecture, i.e., the multi-access techniques employed in the satellite transponder channels to provide communication between a number of geographically scattered ground stations.

Consider the digital communication satellite network depicted in Fig. 1. The network consists of many ground stations and a geostationary satellite. This satellite must be placed in an orbit about 36,000 kilometers away from Earth in order to maintain a stationary position relative to any point on Earth. Each of the satellite transponders acts as a repeater. The ground stations transmit their signals to a transponder at one frequency band. At this transponder the signals are converted into another band, transmitted back to Earth, and delivered to the desired ground station(s).

In most situations, there are many ground stations which must simultaneously reach each other through a satellite transponder. The satellite transponder is a limited resource which must be shared among the ground stations and must meet certain performance criteria

---

<sup>1</sup>This research has been partially supported by a) the U.S. Office of Naval Research under Contract N00014-75-C-0561 and b) National Science Foundation Grant ECS-8017867 with the Department of Operations Research, Stanford University.

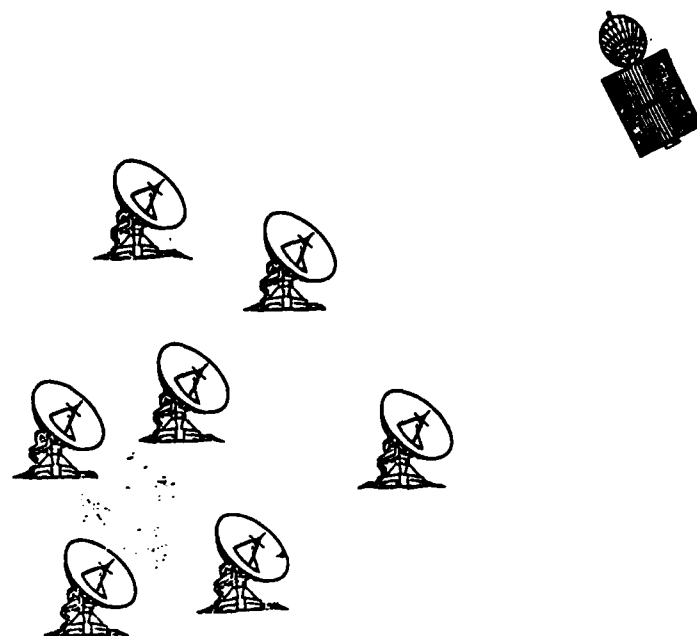


Figure 1. A digital communication satellite network.

concerning the traffic flow between the sources and the destinations. The satellite transponder channel management techniques employed in a multi-access environment consists of two design aspects. One of these concerns the techniques with which we channelize the transponder into smaller units called subchannels. Of interest in this report are the techniques by which the subchannels are derived in time or frequency. These techniques are referred to as Time Division Multiple Access (TDMA) and Frequency Division Multiple Access (FDMA), respectively. We concentrate in this report on the general case of the TDMA Scheme, which reduces to the FDMA Scheme in a special case.

The second design aspect concerns the schemes with which we allocate the subchannels among the ground stations. This allocation can

take place on a Fixed Assigned (Preassigned), Random Access (Contention), or Demand Assigned (Reservation) basis. The case of interest in this report is the fixed assigned scheme - that is, we are concerned with the performance characteristics of a pool of subchannels derived in time or frequency and allocated permanently to the ground stations.

The traffic flow between the ground stations may represent a variety of message types with different constraints. The messages may be the information from a user to a computer, a computer to a user, a telephone conversation, facsimile electronic mail, or it may be video information. The messages are assumed to be comprised of one or more digital data packets that are stored in a memory buffer at the originating ground station. The data packet consists of a fixed number of binary digits.

The memory buffer size at each ground station is finite. Therefore, the messages which arrive after the memory buffer is filled are subject to "blocking," whereby they are lost to the system. Thus, the probability of blocking is one basic performance criterion. Another key measure of performance is the expected delay in the buffer before transmission occurs for messages that are not blocked. Other possible criteria include the probability of exceeding the delay time which these messages can tolerate before reaching the destination.

The buffer behavior of a fixed assigned TDMA system has been analyzed by Chu for Poisson input traffic [1] and batch Poisson input traffic [2] cases. Chu and Konheim [3] have derived the probability generating function of the number of packets in the buffer just before a transmission occurs and the expected delay experienced by the packets.

In their analysis, they have assumed buffers of infinite capacity and have employed the notion of "virtual" message arrival in the expected delay calculation. The results obtained in [3] may serve as a good approximation for practical applications involving finite buffer capacity if the probability of buffer overflow is sufficiently small.

Hayes [4], Spragins[5], Lam [6], and Kosovych [7] also have derived the expected delay experienced by messages under various assumptions for the case of infinite buffer capacity. Various analytical techniques were employed to obtain the steady state movement generating function of the buffer size.

Although assuming infinite buffer capacity considerably simplifies the analysis, it does prevent studying the effect of blocking on the performance of the system. Therefore, this report departs from the investigations described above by making the more realistic assumption of finite buffer capacity. In our approach, we formulate queueing models having a finite queue to represent the system. We then are able to derive such measures as the probability of blocking, the expected number of messages blocked per period, and the expected delay in the buffer for messages that are not blocked. This information would be used to analyze the allocation of subchannels to the incoming traffic and to evaluate the performance of the system.

In our approach we have made use of analytical queueing models rather than simulation models because queueing models can provide a reasonably accurate representation of the actual systems and also furnish mathematical formulas which are easy to compute [8]. The models are examined for cases which have specific applications, so care is taken to make reasonable assumptions.

Before formulating our queueing models in Section 4, we shall first describe the nature of digital traffic flow (Section 2) and of a satellite communication channel (Section 3).

## 2. Characteristics of Digital Traffic Flow

A message which arrives at a ground station to await transmission through the satellite consists of (or is converted into) a series of binary digits (bits). Each message may come from either a single source or a group of sources whose output has been combined (multiplexed) into a single stream of digital traffic. Examples of messages are a telephone call or an inquiry-response between a terminal and a computer.

The messages arrive according to some stochastic process and the length of these messages has some probability distribution. As discussed below, we will approximate such processes with well-known distributions.

Once presented to the ground station, the messages require the transmission medium to transmit them to another ground station within a time constraint. The specification of this time constraint depends highly on the type of message. In Section 2.3, we will describe further the nature of such constraints.

### 2.1. Message Arrival Process

For any given ground station, suppose that messages to be transmitted through the satellite arrive at times  $t_0 < t_1 < t_2 < \dots < t_n$ . (If the message requires conversion into digital form, we assume that the time needed to do this is negligible.) The interarrival time defined by

$\tau = t_n - t_{n-1}$  ( $n \geq 1$ ) is a random variable. We assume that the messages arrive according to a renewal process, so that the interarrival times are independent and identically distributed. This is not always the case in applications because of a time-dependence of the interarrival times, but it should be at least a reasonable approximation for the period of peak usage, which is the particularly interesting case for design purposes.

If the arrivals take place according to a Poisson process, then the interarrival times will be exponentially distributed; that is,

$$\text{Prob} \{ \tau \leq t \} = 1 - e^{-\lambda t}, \text{ for } t \geq 0,$$

where  $\lambda$  is the arrival rate. The probability that  $n$  arrivals occur during time  $t$  is

$$P_n(t) = \frac{\exp(-\lambda t) \cdot (\lambda t)^n}{n!}, \quad n=0,1,2,\dots$$

which is a Poisson distribution.

It has been shown [9] that the assumption of a Poisson process for the arrival of messages has been a reasonable one when the number of users generating these messages is large. Examples are the customers who wish to make a telephone call or the users who send an inquiry from a terminal.

## 2.2. Message Length

Another statistical characteristic of a message is its length, which typically is measured by the number of bits or blocks of bits in

the message. In general, the messages may be either fixed length or variable length. The messages may be segmented into fixed size blocks (packets); in such a case, we will have either a fixed or variable number of packets [10].

Now let the random variable  $Y$  be the message length (expressed as the number of bits). It is usually assumed that  $Y$  has an exponential distribution, so that

$$\text{Prob } \{Y \leq y\} = 1 - e^{-\mu y}, \quad \text{for } y \geq 0$$

where  $1/\mu$  is the expected message length. This assumption has been shown to be valid in many applications [11].

For segmented messages, denote the packet size by  $P$  bits/packet and let the random variable  $L$  be the number of packets in a message, so that

$$L = \langle Y/P \rangle ,$$

where  $\langle x \rangle$  denotes the least integer greater than or equal to  $x$ . It easily follows from the lack-of-memory property of the exponential distribution that  $L$  has a geometric distribution such that

$$\text{Prob } \{L = \ell\} = pq^{\ell-1}, \quad \text{for } \ell = 1, 2, \dots$$

where  $q = e^{-\mu P}$  and  $p = 1 - q$ .

### 2.3. Delay Constraints

Another important characteristic of a message is its associated delay constraint; that is, when a message to be transmitted arrives at a station, it should be transmitted and received at its destination within some time limit. We define the total delay as the elapsed time from when a message arrives to be transmitted until it is received at its destination. This elapsed time includes very small amounts of time needed for possible digitization, packetization, depacketization and any other processing at the ground stations (so we hereafter assume that these times are negligible). Thus, the total delay essentially is just the delay in the buffer plus the time required (see Section 3) for the message to travel to the satellite and back to its destination.

The delay constraint specification is very important since it determines, along with the other digital traffic characteristics, what system architecture should be considered for service. We will elaborate on this matter in the next section. The most common methods for specifying the delay constraint are:

- Maximum delay
- Average delay

The maximum delay for service is sometimes used as a constraint by giving the delay which can be exceeded by no more than a given percentage of messages. The average delay constraint, expressed as an upper bound on expected delay, is usually specified for message types which do not require an exact amount of tolerable delay.

Based on the type of application and the nature of the digital traffic (which may be data, voice or image), the user may describe the constraint in any of the above forms.

### 3. Characteristics of a Satellite Channel

The transmission medium which provides service for digital traffic flow is the satellite channel. Its limited capacity is divided among the ground stations by dividing the channel into time or frequency bands (subchannels). The subchannels are allocated either permanently or temporarily to the ground stations.

The need to characterize the service quality is evident, since it is important to evaluate how effectively the channel is being allocated to match the digital traffic requirements. Some properties of the satellite channel that enhance its service quality are its multi-access property and its ability to broadcast. One limitation is that, due to the large distance between a satellite and Earth, it takes about 0.27 seconds for the electromagnetic waves to travel from Earth to the satellite and back to Earth. This delay introduces a bound on the service that can be provided to incoming traffic. The important channel design parameters will be described in the following subsections.

#### 3.1. Subchannels

Subchannels are obtained by partitioning the satellite transponder channel into frequency or time slots, corresponding to FDMA and TDMA, respectively. Consider a time-shared system with a number of ground stations which transmit their messages in a sequence of non-overlapping time slots that use the entire transponder channel. The entire time needed to fill all of these slots once is called the frame time. Figure 2 shows the TDMA structure where the frame time  $T$  is divided into  $K$

equal slots, one or more of which are assigned to each ground station.

The messages are assumed to be segmented into one or more packets of slot size, so that one packet of  $P$  bits can be packed in each slot. The value of  $P$  is considered an important aspect of the design. Reference [13] discusses the selection of an optimal block size for packet transmission.

The channel transmission rate,

$$C = K P/T ,$$

is the most important channel parameter and is assumed to be constant over time.

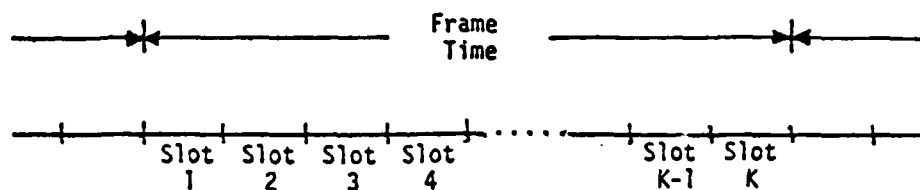


Figure 2. A time shared channel.

### 3.2. Traffic Intensity

For a ground station, its traffic intensity,  $a$ , is defined as

$$a = E[Y]/E[\tau] = \lambda/\mu$$

where  $E[\tau] = 1/\lambda$  is the expected interarrival time of messages, and  $E[Y] = 1/\mu$  is the expected message length. The unit of traffic intensity is the Erlang. Normalized traffic intensity is defined to be the traffic intensity per slot; i.e., if there are  $S$  slots permanently assigned to the source, then its normalized traffic intensity is

$$\rho = a/S.$$

The traffic intensity or normalized traffic intensity for the entire satellite channel also can be defined in an analogous manner. In this case,  $S$  would correspond to  $K$  of Section 3.1, and the normalized traffic intensity would more commonly be referred to as the channel utilization.

### 3.3. Scheduling Policy

Another important aspect of service is the scheduling policy for processing the messages. The most common scheduling policies are First-In-First-Out (FIFO) and Random. In FIFO the messages arriving at each ground station are processed by the channel according to their order of arrival, whereas in random service the messages are processed on a completely random basis. The FIFO policy normally would be an appropriate one when the messages have the same delay time constraints. If the incoming messages have different delay constraints, then those

requiring less delay take priority over the long-delay ones. This is referred to as service priority and, in practice, different strategies for handling various priority-class messages are adopted.

#### 4. Models for Analyzing the Performance of a Satellite Channel

In this section we introduce a basic type of model for the Fixed Assigned allocation scheme and derive the key measures of performance, namely, expected delay and blocking probability in steady state. The model can be adapted to include other allocation schemes. The results of the analysis generate appropriate vehicles for analyzing the performance of many practical systems.

##### 4.1. Model Formulation

Consider  $N$  ground stations sharing a channel with a transmission rate (capacity) of  $C$  bits/sec. Assume that the frame time  $T$  is divided into fixed slots. Let us focus on a typical ground station which can transmit packets in  $S$  consecutive allocated slots in each frame. These slots are assigned to this ground station and no other user has access to them. The messages arriving for transmission from this ground station are stored (if there is room) in its finite capacity buffer and then transmitted in these slots, as shown in Fig. 3. For purposes of analyzing the performance of the system in transmitting messages from this typical ground station, the messages and transmissions from the other  $(N-1)$  ground stations can and will be ignored.

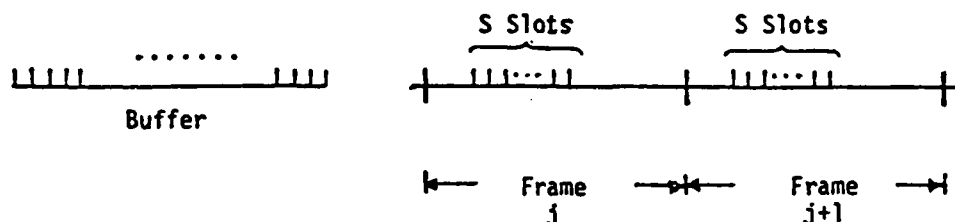


Figure 3. The Fixed Assigned allocation scheme.

Assumptions:

- Arriving messages occur according to a Poisson process at a mean rate of  $\lambda$  messages/sec.
- Each message is segmented into one or more packets of  $P$  bits per packet.
- The buffer has a fixed capacity for storing a maximum of  $M$  packets. Any message (or portion thereof) will be accepted into storage as long as there is room remaining. (If there is room for only a portion of a message, we assume as an approximation that this portion is accepted and treated like an entire message.) Otherwise, it will be blocked and will leave the system.
- Messages are transmitted (one packet per slot) according to a FIFO (First-In-First-Out) policy.

- $S$  is small relative to the total number of slots in a frame so that, as an approximation, the transmission of messages in these  $S$  slots is considered to occur simultaneously. Without loss of generality, these slots are assumed to be the first  $S$  slots in each frame, so that their transmissions are considered to occur at the instant each frame begins, with a time lapse of  $T$  between their transmissions.

In the terminology of queueing theory, this is a batch service model with a single server, a finite queue, a Poisson input (Sec. 4.2) or compound Poisson input (Sec. 4.3), and constant service times  $T$ . The "customers" are the packets, the "queue" is the buffer, and the "server" is the channel, where the server is considered to be tied up in service during both the transmission of the  $S$  packets and the "recovery time" until the server is ready to begin the  $S$  slots in the next frame. However, a key difference from the standard models of queueing theory is that the server here always is busy, since it will become tied up in "serving" the next batch of customers even when there are no customers in this batch.

In the following two sections we examine the performance of this system for single-packet and multi-packet messages.

#### 4.2. The Model for Single-Packet Messages

Let us assume that the messages have lengths shorter than the slot sizes so that each message consists of just one packet that will be stored in one buffer unit and will be transmitted in only one slot. As indicated above (Sec. 4.1),  $S$  slots are allocated in each frame to

transmitting the  $S$  messages at the head of the buffer (or all of the messages if there are less than  $S$  in the buffer).

Consider the imbedded Markov chain obtained by observing the system at just those instants (epochs) where a frame begins. Let  $a_n$  denote the probability of  $n$  arrivals between epochs (i.e., during a frame), that is,

$$a_n = \frac{(\lambda T)^n \cdot \exp(-\lambda T)}{n!}, \text{ for } n=0,1,2,\dots \quad (1)$$

Letting the state of the system be the number of messages in the buffer, excluding those which just were removed for transmission, we may write the following transition probability matrix.

$$\bar{P} = \begin{matrix} & \begin{matrix} 0 & 1 & \dots & m-1 & m \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ \vdots \\ S \\ S+1 \\ \vdots \\ m \end{matrix} & \begin{bmatrix} \sum_{n=0}^S a_n & a_{S+1} & \dots & a_{M-1} & \sum_{n=M}^{\infty} a_n \\ \sum_{n=0}^{S-1} a_n & a_S & \dots & a_{M-2} & \sum_{n=M-1}^{\infty} a_n \\ \vdots & \vdots & & \vdots & \vdots \\ a_0 & a_1 & & \vdots & \vdots \\ 0 & a_0 & & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & a_{S-1} & \sum_{n=S}^{\infty} a_n \end{bmatrix} \end{matrix} \quad (2)$$

where  $m = M - S$ .

Therefore, the state probabilities  $\{\pi_j, j=0,1,\dots,m\}$  are given by

$$\pi = \pi \bar{P} \quad \text{where} \quad \pi = [\pi_0, \pi_1, \dots, \pi_m]$$

or

$$\pi_j = \sum_{i=0}^m \pi_i P_{ij}, \quad \text{for } j=0,1,\dots,m.$$

A message is blocked (rejected) if the Markov chain was in state  $j$  and then there were at least  $(M - j)$  preceding arrivals of messages before the next frame starts. Thus, letting  $R$  be the number of such messages rejected in one frame, the expected number of messages blocked per period (frame) is

$$E(R) = \sum_{j=0}^m \pi_j \sum_{n=0}^{\infty} n a_{M-j+n} \quad (3)$$

where the second summation is written in terms of a finite number of elements as

$$\sum_{n=0}^{\infty} n a_{M-j+n} = \lambda T - \sum_{n=0}^{M-j-1} n a_n - (M-j)(1 - \sum_{n=0}^{M-j-1} a_n) \quad (4)$$

The probability that a message will be blocked (rejected) is given by

$$P_B = E(R)/\lambda T. \quad (5)$$

Now we proceed to calculate  $W_q$ , which is defined as the expected delay in the buffer incurred by those messages which will be transmitted, i.e., those which are not blocked. (In the terminology of queueing theory,  $W_q$  is called the expected waiting time in the queue.) Let the random variable  $W$  be the delay in the buffer incurred by a randomly selected arriving message (regardless of whether it is blocked or not). Note that  $W > 0$  implies that the message is not blocked, whereas  $W = 0$  implies that the message is blocked (with probability 1) since the probability that a message arrives at exactly the same instant that a frame begins (transmission occurs) is zero. Therefore,

$$\begin{aligned} E(W) &= E(W|W > 0)(1 - P_B) + E(W|W = 0)P_B \\ &= W_q (1 - P_B) , \end{aligned} \tag{6}$$

so that calculating  $E(W)$ , along with  $P_B$  from Eq. (5), will immediately yield  $W_q$ . Hence, we now will focus on deriving  $E(W)$ .

For the random arrival under consideration, let  $E_{in}$  be the event that this arrival meets the conditions: (1) finds  $i$  messages left behind in the buffer from the last epoch (start of new frame), and (2) is one of  $n$  arrivals in this frame. Therefore, letting  $P(E_{in})$  denote the probability that the event  $E_{in}$  occurs for the random arrival,

$$E(W) = \sum_{i=0}^m \sum_{n=1}^{\infty} P(E_{in}) E(W|E_{in}) . \tag{7}$$

To find  $P(E_{in})$ , note that the conditional probability given  $i$  that a given random arrival is one of  $n$  arrivals in this frame must be weighted by  $n$ , so that

$$P(E_{in}) = \pi_i \frac{n a_n}{\sum_{j=1}^{\infty} j a_j} . \quad (8)$$

(This second term also can be interpreted as the expected fraction of arrivals that are in a group of  $n$  arrivals.) The problem now is reduced to finding  $E(w|E_{in})$ . To do this, we distinguish two cases:

Case 1:  $i + n \leq M$

For this case, all  $n$  arrivals that occur in the frame interval  $[0, T]$  will wait. Therefore, a key observation (see [14]) is that, with a Poisson input, the time of the  $j^{\text{th}}$  arrival out of  $n$  over a unit interval has a Beta distribution with parameters  $p = j$ ,  $q = n - j + 1$ . Thus, the time of this arrival over the interval  $[0, T]$  has

$$\text{mean} = \frac{p}{p+q} T = \frac{j}{n+1} T.$$

It then follows that

$$\begin{aligned} E(\text{cumulative waiting by these } n \text{ until } T) &= \frac{1}{n+1} \sum_{j=1}^n (n+1-j)T \\ &= \frac{1}{n+1} \frac{n(n+1)}{2} T = n \frac{T}{2} \end{aligned}$$

where this expression excludes any additional waiting in the buffer

that may be incurred by these  $n$  messages in subsequent frames. (This result also can be obtained directly from the fact that the  $n$  arrivals are uniformly distributed over  $[0, T]$ .) Adding in the waiting in subsequent frames, we then have

$$E(\text{total cumulative waiting for all of these } n \text{ arrivals}) \\ = n T/2 + \min \{n, (n + i - S)^+\} T + \min \{n, (n + i - 2S)^+\} T + \dots \quad (9)$$

where  $(x)^+ = \max \{0, x\}$ .

Therefore, the average waiting for each of these arrivals will be the expression given in Eq. (9) divided by  $n$ ; i.e.,

$$E(w|E_{in}) = T/2 + \sum_{k=1}^{\infty} \min\{n, (n + i - Sk)^+\} T/n \quad (10)$$

when  $i + n \leq M$ . Note that the terms in the summation become zero as soon as  $Sk \geq n + i$ .

Case 2:  $i + n > M$

For this case, the  $i$  messages already in the buffer plus the  $n$  arrivals during the current frame exceed the waiting room in the buffer, so  $(M-i-n)$  messages are blocked and  $\bar{n} = (M-i)$  new arrivals wait. For these  $\bar{n}$  new arrivals who wait we have:

$$E(\text{cumulative waiting by these } \bar{n} \text{ until } T) = \frac{1}{\bar{n}+1} \sum_{j=1}^{\bar{n}} (n+1-j)T$$

$$\begin{aligned}
&= \frac{1}{n+1} \left[ \frac{n(n+1)}{2} T - \frac{(n-\bar{n})(n-\bar{n}+1)}{2} T \right] \\
&= \left[ n - \frac{(n-\bar{n})(n-\bar{n}+1)}{n+1} \right] T/2 .
\end{aligned}$$

$$\begin{aligned}
\text{Therefore, } E(w|E_{in}) &= \frac{1}{n} \left[ n - \frac{(n-\bar{n})(n-\bar{n}+1)}{n+1} \right] \frac{T}{2} \\
&\quad + \frac{1}{n} \sum_{k=1}^{\infty} \min\{\bar{n}, (\bar{n}+i-S_k)^+\} T
\end{aligned} \tag{11}$$

when  $i + n > M$ . Note that the terms in this summation become zero as soon as  $S_k \geq \bar{n}+i$ .

#### 4.3. The Model for Multi-Packet Messages

We now assume that some or all of the messages have lengths longer than the slot sizes. In this case we segment the message length into slot-size packets. Recall that the random variable  $L$  is the number of packets in a message; denote its distribution by  $g(\ell)$ . Thus, it is being assumed that messages arrive according to a Poisson process where each message contains  $L$  packets. The combination of these two random factors will result in a compound Poisson distribution; that is, the probability of  $j$  packet arrivals,  $a_j$ , in an interval of fixed length is obtained by convolving  $g(\ell)$  a random number of times according to the Poisson distribution.

Once again, an imbedded Markov chain can be constructed just as described in Section 4.2, where the state of the system now is the number of packets (rather than messages) in the buffer. Let us make the

simplifying assumption that if the buffer has room for just some of the packets in an arriving message, these packets are accepted into the buffer (and considered a complete message) whereas the rest of the packets are blocked. Then the transition probability matrix still is given by Eq. (2), except for the difference in calculating  $a_j$  described in the preceding paragraph, so the steady-state probabilities  $\pi_j$  also are obtained the same way. However, the problem of finding the expected message delay is somewhat different now because a message is considered to be delayed as long as any of its packets are still delayed.

Except for taking this difference into account, we can follow the same steps used in Sec. 4.2 to derive  $W_q$ , the expected delay for messages which are not blocked. To find  $E(R)$ , the expected number of messages rejected (blocked) in one frame, let the random variable  $R_j$  be the number of messages rejected given that  $j$  packets were left behind in the buffer at the beginning of the frame. Also let  $G^{(i)}(t)$  be the probability that the number of packets in  $i$  messages is less than or equal to  $t$ , which would be calculated by taking the  $i$ -fold convolution of  $g(\ell)$ . Finally, let the random variable  $N$  be the number of messages arriving in a frame, which is assumed to have a Poisson distribution with parameter  $\lambda T$ , so that

$$\text{Prob } \{N = k\} = \frac{(\lambda T)^k \exp(-\lambda T)}{k!}, \quad \text{for } k=0,1,2,\dots \quad (12)$$

Therefore,

$$\text{Prob } \{R_j = n\} = \sum_{i=1}^{M-j} \text{Prob } \{N = n + i\} [G^{(i-1)}(M-j-1) - G^{(i)}(M-j-1)] \quad (13)$$

and

$$E(R) = \sum_{j=0}^m \pi_j \sum_{n=1}^{\infty} n \text{ Prob } \{R_j = n\} \quad (14)$$

The resulting probability that a message will be blocked is

$$P_B = E(R)/\lambda T \quad (15)$$

Given  $P_B$ ,  $W_q$  again can be obtained directly from  $E(w)$  by using Eq. (6), so we now will derive  $E(w)$ .

For the randomly selected arriving message under consideration, let  $E_{ijnpq}$  be the event that this message meets the following conditions:

- (1) finds  $i$  packets left behind in the buffer at the beginning of the frame,
- (2) is the  $j^{\text{th}}$  out of  $n$  arriving messages in the current frame,
- (3) the total number of packets in the first  $(j-1)$  messages is  $p$ ,  
and
- (4) the number of packets in this  $j^{\text{th}}$  message is  $q$ .

Also, let  $P(E_{ijnpq})$  be the probability that the event  $E_{ijnpq}$  occurs for this message.

Note that  $E(w|E_{ijnpq}) = 0$  automatically (due to blocking) if either  $j > M-i$  or  $p > M-i-1$ . Therefore,

$$E(w) = \sum_{i=0}^m \sum_{j=1}^{M-i} \sum_{n=j}^{\infty} \sum_{p=j-1}^{M-i-1} \sum_{q=1}^{\infty} P(E_{ijnpq}) E(w|E_{ijnpq}) \quad (16)$$

Let  $r = \min \{p + q, M - i\}$ , and let  $g^{(j-1)}(p)$  be the probability that the first  $(j-1)$  messages contain  $p$  packets (as obtained from the  $(j-1)$  - fold convolution of  $g(x)$ ). As observed in Section 4.2, the time of the  $j^{\text{th}}$  arrival out of  $n$  over a unit interval has a Beta distribution. It then follows that

$$E(W|E_{ijnpq}) = \frac{n+1-j}{n+1} T + < \frac{i+r-S}{S} > T. \quad (17)$$

Proceeding as for Eq. (8), it also follows that

$$P(E_{ijnpq}) = \pi_i \frac{1}{n} \frac{n P\{N=n\}}{\sum_{k=1}^{\infty} k P\{N=n\}} g^{(j-1)}(p) g(q) \quad (18)$$

Combining Eqs. (6) and (12) to (18) now yields  $W_q$ .

In a subsequent report, we will discuss the application of the above results (Sections 4.2 and 4.3) to several practical cases, as well as present and analyze some numerical results.

## REFERENCES

- [1] Chu, W. W., "Buffer Behavior for Poisson Arrival and Multiple Synchronous Constant Outputs," IEEE Transactions on Computers, Vol. C-19, June, 1970, pp. 530-534.
- [2] Chu, W. W., "Buffer Behavior for Batch Poisson Arrivals and Single Constant Output," IEEE Transactions on Communications Technology, Vol. COM-18, October, 1970, pp. 613-618.
- [3] Chu, W. W. and A. G. Konheim, "On the Analysis and Modeling of a Class of Computer Communication Systems," IEEE Transactions on Communications, Vol. COM-20, No. 3, June, 1972, pp. 645-660.
- [4] Hayes, J. F., "Performance Models of an Experimental Computer Communication Network," Bell System Technical Journal, Vol. 63, February, 1974, pp. 225-259.
- [5] Spragins, J., "Simple Derivation of Queueing Formulas for Loop System," IEEE Transactions on Communications, Vol. COM-25, April, 1977, pp. 446-448.
- [6] Lam, S. S., "Delay Analysis of a Packet-Switched TDMA System," National Telecommunication Conference Proceedings, 1976, pp. 16.3-1 - 16.3-6.
- [7] Kosovych, O. A., "Fixed Assignment Access Technique," IEEE Transactions on Communications, Vol. COM-26, No. 9, September, 1978, pp. 1370-1376.
- [8] Spragins, J., "Analytical Queueing Models," IEEE Computer Magazine, Vol. 13, No. 14, April, 1980, pp. 9-11.
- [9] Beckmann, P., Introduction to Elementary Queueing Theory and Telephone Traffic, The Golem Press, Boulder, Colorado, 1968, pp. 30-33.
- [10] Chu, W. W., "Design Considerations of Statistical Multiplexors," Proceedings of the ACM Symposium on Problems in the Optimization of Data Communications Systems, Pine Mountain, Georgia, October 1969, pp. 35-59.
- [11] Fuchs, E. and P. E. Jackson, "Estimates of Distributions of Random Variables for Certain Computer Communications Traffic Models," Proceedings of the ACM Symposium on Problems in the Optimization of Data Communications Systems, Pine Mountain, Georgia, October 1969, pp. 206-230.

- [12] Occigross, B., I. Gitman, N. Hsieh, and H. Frank, "Performance Analysis of Integrated Switching Communications Systems," National Telecommunication Conference, 1977, pp. 12: 4-1 - 4-13.
- [13] Wolman, E., "A Fixed Optimum Cell Size for Records of Various Length," ACM Journal, Vol. 12, No. 1, January 1965, pp. 53-70.
- [14] Karlin, S. and H. M. Taylor, A First Course in Stochastic Processes, 2nd ed., Academic Press, New York, NY, 1975.

